DOCUMENT RESUME

ED 334 265 TM 016 890

AUTHOR Hambleton, Ronald K.; Murphy, Edward
TITLE A Psychometric Perspective on Authentic Measurement.
PUB DATE 91
NOTE 33p.
PUB TYPE Information Analyses (070) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cognitive Tests; Comparative Analysis; Elementary Secondary Education; Literature Reviews; *Multiple Choice Tests; *Objective Tests; *Psychometrics; *Testing Problems; Test Use
IDENTIFIERS *Authentic Assessment; Higher Order Learning

ABSTRACT

Authentic measurement has become an important topic recently in educational testing. Advocates of authentic measurement feel that objective tests, multiple-choice tests in particular, cannot meet the demands required of today's tests and should be replaced by tests that can be closely matched to instruction and can assess higher-order cognitive skills. This paper addresses the validity of several criticisms of objective tests and, where appropriate, considers the viability of some alternatives. The four criticisms of objective tests that are considered contain arguments that such tests foster a one-right answer mentality, narrow the curriculum, focus on discrete skills, and under-represent the performance of students from low socioeconomic backgrounds. It is contended that the evidence against multiple-choice tests is not nearly as strong as has been claimed. It is not clear whether authentic measurements are always better. Substantially more research into the strengths and weaknesses of various item formats for meeting particular measurement needs should be conducted. A 49-item list of references is included. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334265

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RONALD K. HAMBLETON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Psychometric Perspective on Authentic Measurement

Ronald K. Hambleton

University of Massachusetts at Amherst

Edward Murphy

National Evaluation Systems

Running Head: Authentic Measurement

TM 016 890

BEST COPY AVAILABLE

A Psychometric Perspective on Authentic Measurement

Abstract

Authentic measurement has become an important topic recently in educational testing. Advocates of authentic measurement feel that objective tests, and, in particular, multiple-choice tests, cannot meet the demands required of today's tests and should be replaced by tests that can be closely matched to instruction and can assess higher-order cognitive skills. The purpose of this paper is to address the validity of several of the popular criticisms of objective tests, and, where appropriate, to consider the viability of some of the alternatives. The four criticisms of objective tests that are considered in the paper are (1) fostering a one-right answer mentality, (2) narrowing the curriculum, (3) focusing on discrete skills, and (4) under-representing the performance of low-SES examinees. Our review suggests that the evidence against multiple-choice tests is not nearly as strong as has been claimed; it remains to be proven whether authentic measurements are always better; and substantially more research as to the strengths and weaknesses of various item formats for meeting particular measurement needs should be carried out.

A PSYCHOMETRIC PERSPECTIVE ON AUTHENTIC MEASUREMENT[1]

Ronald K. Hambleton
University of Massachusetts at Amherst

and

Edward Murphy[2]
National Evaluation Systems

Every year, it seems, a new important issue is introduced into the national debate about educational assessment. "Truth in testing," "Minimum Competency Testing," "The Lake Wobegon effect," and "the Golden Rule settlement" are four recent entries into the debate that have influenced testing practices considerably. The new issue in 1990 was "authentic measurement," and its impact on testing practices in 1991 and beyond could be immense. At the annual meeting of NCME in Boston in 1990, only a few papers addressed the topic. This was followed by the National Centers Competition, sponsored by OERI, in which bidders were strongly advised to address the national need for authentic measurement in their five-year research plans. State departments of education, too, have reflected some of the same OERI language in their recent testing RFPs. In 1990, ETS sponsored a conference which addressed the need for an expansion in the assessment methods that are used. And, journals such as Educational Leadership devoted a whole issue to the topic. At the 1991 meeting of NCME

---

[1]Laboratory of Psychometric and Evaluative Research Report No. 214. Amherst, MA.

[2]Currently a doctoral student at the University of Massachusetts at Amherst.

in Chicago, over 30% of the program, and the President's address (Carlson, 1991) were focused on the topic. There is little doubt, therefore, about the importance and current interest level among educators and testing specialists in authentic measurement.

Authentic measurement is historically what measurement specialists have called "performance testing," with writing assessments, Red Cross swimming tests, and the driving tests being three well-known examples. The goal of authentic measurement appears to be to make assessment more closely resemble actual learning tasks and to permit the assessment of higher-order cognitive skills such as problem-solving, critical thinking, and reasoning. Of special importance to many advocates of authentic measurement is the capability for assessment to permit multiple correct answers and/or multiple acceptable methods to solve a problem or complete a task.

As with other recent national issues such as truth in testing and inflated norms (called "the Lake Wobegon effect"), persons working outside the measurement community are at the forefront of the debate and are exerting considerable pressure to bring about changes in educational testing practices. Curriculum specialists, policy-makers, and teachers are among the leading advocacy groups. These groups argue that more valid measurements would result and more acceptance of testing would occur if objective test formats, notably the multiple-choice format, were de-emphasized and, in their place, oral reports, exhibitions, projects, portfolios, performance assessments, writing samples, observations, self- and peer-assessments, reviews, etc., were substituted.

Advocates for authentic assessment in education appear to want to bring testing methods more in line with instruction. They want assessments to closely approximate what it is students should know and be able to do: complete a science experiment, write a persuasive essay, prepare a report on farming in New England, deliver a speech, etc. At a philosophical level, it is hard to disagree. Assessment methods are needed to measure these and other skills, and objective testing methods will not always suffice. But, it would be incorrect to argue that authentic assessment in education is new. The name is new but the ideas underlying authentic measurements are reflected in performance testing, which has had a long history in the field of psychometric methods.

At a practical level, obtaining authentic measurements is going to be a major challenge. First, few educational measurement specialists have had very much experience in constructing and using performance tests. (Some of the best work has been done by persons working in the Armed Services and in industry.) But, if educational measurement specialists are not well-informed, most classroom teachers and administrators would know even less. In education, experience is limited to writing assessments, some performance testing in the credentialing examination area, and a few other isolated areas in science and mathematics. And, even these relatively few performance tests are not without their critics. Second, performance tests are, in many instances, going to take a lot more time to construct, to administer, and to score than objective tests. And, principles laid out in the AERA, APA, and NCME Standards for Educational and Psychological Testing for standardization, reliability, and validity will apply equally to

performance tests as apply to objective tests. Clearly, then, the challenges of successfully obtaining authentic measurements will be great.

What is the answer to the challenges posed by authentic measurement? Certainly more varied assessment methods need to be used in the future to assess a variety of important performances and products, such as writing skills, oral communication skills, etc. Therefore, the two problems described above, shortage of trained personnel and the difficulties associated with test development, scoring, and validity assessment will need to be overcome.

But what about objective tests, especially multiple-choice tests? Should they play any role in future assessments beyond their use in the assessment of lower-level skills? Advocates of authentic measurement have been fairly critical of objective test formats (e.g., Wiggins, 1990), but how valid are the criticisms? In turn, some measurement specialists have argued that more can be done with objective formats than critics have acknowledged. And, if objective formats can meet some of the expectations for assessment in the 1990s, then more time and resources would be available to tackle those difficult to measure skills that cannot be handled with objective formats. The main purpose of this paper, then, is to address the validity of several of the popular criticisms of objective tests and, where appropriate, to consider the viability of some of the alternatives.

Popular Criticisms of Objective Tests

Our literature review of authentic assessment uncovered a number of criticisms of objective tests:

1. Fostering a One-Right-Answer Mentality,
2. Narrowing the Curriculum,
3. Focusing on Discrete Skills, and
4. Under-representing the Performance of Lower-SES Examinees.

In what follows, the nature of each criticism will be addressed, and then the merits of alternative methods of assessment will be considered. In a second paper (Hambleton & Murphy, 1991), four additional criticisms of objective tests are considered: (1) Lacking the capability to measure higher order cognitive skills, (2) focusing on the product at the expense of process, (3) selecting answers rather than creating them, and (4) lacking construct validity.

1. Fostering a One-Right-Answer Mentality

Objective tests are sometimes criticized not only for what they do not do (i.e., assess higher-order thinking, test knowledge of process, etc.) but also for what they supposedly *do* to examinees. Multiple-choice test items, in particular, are criticized for their potentially limiting effects on the thinking processes of students. Because only one answer choice in a multiple-choice item can be correct, it has been argued that excessive exposure to such items may engender a simplistic view of the world (e.g., Nader, 1987). It has been conjectured that examinees may develop a "one-right-answer mentality," expecting that events and problems in the real world will mirror the artificial setups in multiple-choice items by yielding to a single correct solution. This perception may, over time, produce persons who are at a disadvantage in coping creatively and

effectively with the subtleties and nuances of real-life problems (Marzano, et al., 1988).

Discussion of the criticisms. Multiple-choice items do promise examinees that they will find one correct or best response among the choices offered. But there is a difference between the single correct or best answer from among four or five stated alternatives and the unique correct answer from among the infinite universe of potential responses. And it is the former, not the latter, that multiple-choice items explicitly offer to examinees.

Many multiple-choice items do pose problems for which there is only one right answer; but this reflects reality. For some problems, there is only one right answer. If, for example, a person needs to figure out how to attach the wires in an electric light fixture, there may be several ways he or she can address the problem, but there is in fact only one right answer. For assessing knowledge in this type of situation the "simplistic" approach of a multiple-choice question seems to fit and to be both harmless and useful.

On the other hand, there are many problems and questions in real life that have multiple solutions; so, too, multiple-choice items can be written that do not pretend to offer the only conceivable right answer to the questions they pose. For example, consider this item stem: "Which of the following characteristics of U.S. society in the early nineteenth century was a major cause of the Civil War?" No claim is made that there is only one right answer to the question; in fact, the contrary is implied in the wording of the stem. The only claim made by the item is that there is only

one right answer printed on the page below the item. And this sort of item is by no means rare: a large number of multiple-choice test items are "best answer given" items rather than "one right answer" items.

Even if multiple-choice items did indeed present exclusively one-right-answer situations, students are exposed to many different question types, in writing and orally, throughout their schooling, including a great number that would tend to counteract the effects of "one-right-answerism." Also, the supposed saliency and power of the multiple-choice test (or any assessment) in forming the value structures of students, exposed as they are to such a wealth of more vibrant stimuli, is unproven.

Alternative assessment approaches. There is substantial merit in using other item formats than multiple-choice in many situations, including large-scale assessments and credentialing exams. If policies and budgets permit, short-answer items, essay-type items, or other sorts of supply-type items may be used. However, as discussed above, such item types present problems of scoring reliability, cost, test length, and turnaround time for grading that must be satisfactorily dealt with by users.

If such problems prove intractable, there are machine-scorable item types other than multiple-choice items that are designed to avoid the one-right-answer quandary. One approach is simply to create multiple-choice items that contain more than one correct response (e.g., the multiple true-false format), with examinees receiving full credit for selecting every correct response presented. Designing a reasonable method for assigning partial credit to partially correct responses of several sorts is one of the difficulties with this approach, but it does illustrate that there are

ways to address the many-right-answers situation even in the most traditional item format. See, for example, the work of Masters and Wright (1984) for IRT measurement models that can handle the type of data available with this item format.

The multiple true-false item is not well-known in objective testing, though it has the important advantage of permitting multiple correct answers. In this format, examinees are presented with a situation (usually described in writing, but sometimes presented via videotape or videodisc) and are asked to select any number of correct responses to it. Thus, for example, an examinee in a medical context may be presented with a set of symptoms and a list of possible treatment options, and may be asked to select (Yes/No) all treatment options that would be potentially beneficial in treating the symptoms presented by the patient, or, conversely, all options that might be harmful. This format fosters careful deliberation in the examinee as he or she selects from a given list a (usually) unspecified number of viable options and rejects a similarly unspecified number of inappropriate ones. While maintaining the benefits of machine scorability, such a format is certainly a far step from the one-right-answer approach.

In sum, it is clear that a one-right-answer slant can easily be avoided by a judicious combination of traditional multiple-choice items and other formats, both machine-scorable and not. In fact, the tradeoff between the validity of supply-type items and the reliability of selection-type formats is often a false dilemma. Test users need not sacrifice the proven reliability of traditional item formats for a perceived increment in the validity of their assessments; it is often possible to combine item

types to meet both ends. Well-written objective items can approach the validity of more experimental or subjective item types; carefully constructed and scored open-ended items can approach the reliability of more objective item types. Used in combination, each type can enhance the benefits and compensate for the limitations of the other.

2. Narrowing the Curriculum

A strong objection to current objective testing is that the very limited parameters of the content and approaches embodied in the typical test - again, the multiple-choice test especially - too often define the content and approaches of instruction in the classroom. By providing policymakers, lawmakers, parents, and the general public with an all-too-simple yardstick for measuring the quality of instruction, large-scale tests become powerful forces in the curricular decision-making of teachers and administrators, and consequently in the learning of students. And if the tests are off-target in their content, outmoded in their approaches, or inaccurate in their assumptions, these deficiencies can be translated into the instruction that is conveyed to students because test content, approaches, and assumptions can influence course content, approaches, and assumptions (Frederiksen, 1984; Shepard, 1989).

Discussion of the criticisms. Standardized tests can have, and have had, a significant effect on instruction; some educators have even applauded this fact and refer to it as "measurement-driven instruction" (Popham, 1987). The reasons for the strong influence of testing on instruction are not hard to find. When general satisfaction with a school can be obtained with successful performance on a standardized test, and

when widespread grief follows poor performance, it is no wonder that some school officials and teachers focus their attention on finding out what the tests will be like and what they will cover, and then teach the format and content of the tests until they are adequately mastered, even overmastered. If in this process of teaching to the test other, richer content and approaches are necessarily given short shrift because of instructional time limitations, the consequences are far less perceptible, immediate, and punitive than they would be if the neglect were on what is covered on the tests. It is human nature to pursue behaviors that evoke positive outcomes and avert painful ones, and not to worry about those that apparently have neutral consequences.

However, while the fact that many teachers teach to the test is undeniable, there is nothing special about objective tests or multiple-choice item formats that evokes this self-protective reaction from teachers and administrators. Any kind of test that produces results that can be compared to results from other schools will become the focus of educators' attention and ameliorative efforts. In fact, it is the acceptance of the virtual universality of the fact that tests influence instruction that causes advocates of authentic measurement to insist that special care be given to the nature of the test: If teaching to the test is inevitable, they argue, the test must be worth teaching to (Frederiksen & Collins, 1989; Marzano, et al., 1988; Nickerson, 1989; Wiggins, 1989). This is an important point and was an argument used to support the increased use of criterion-referenced testing programs in the 1970s.

It would be desirable if educators would have enough confidence in their own teaching abilities and their students' learning capacities to proceed with their classes as if there were no tests to be given (Mehrens & Kaminski, 1989). However, the stakes for teachers and administrators are high in most schools, and so such thoughts are probably unrealistic. Instead, it seems more reasonable to focus on constructing/selecting tests for schools that are of such soundness, validity, and reliability that it would be clearly acceptable if the curriculum were in fact modeled on their content and format.

Alternative assessment approaches. We agree with advocates of authentic assessment that worthy assessment instruments are desirable because of their effects on instruction (as well as for their measurement purposes). But, unless objective item formats (in contrast to norm-referenced achievement tests) are proven to be useless or harmful, they can be used effectively, and sometimes in combination with less traditional (i.e., authentic) assessment approaches (ones that have been shown by research to be sound both instructionally and psychometrically) to meet the needs and concerns of both testing and instructional professionals.

3. Focusing on Discrete Skills

Another common criticism of objective tests, especially criterion-referenced tests, is that they do not adequately reflect the emerging sense of interrelatedness and holism that is affecting many of our formerly separatist, atomistic disciplines. The main criticism is that just at the time when grand syntheses are overtaking such disciplines as physics, chemistry, biology, mathematics, and computer science (for example),

objective tests (especially many criterion-referenced tests) persist in dividing up the universe into small, discrete components (Marzano, et al., 1988). The result of such an analytical approach, according to critics, is an incoherent, disjointed conception of a discipline that does not reflect the ways that its practitioners think and act when they are practicing it. For example, consider a French test that focuses on the present tense in one objective, the indirect object in another, and the comprehension of everyday dialogue in a third. Examinees who perform well on discrete objectives (even every discrete objective) are not necessarily able to use French competently, and examinees who do poorly on each objective may not be poor users of the language (Commission on Reading, National Academy of Education, 1985). Clearly, there is a problem here of construct validity.

Discussion of the criticisms. There has been a definite tendency on the part of test developers to divide big skills or competencies into smaller ones for the purposes of testing. But this tendency has not been arbitrary in the negative sense of that word. Three explanations for this tendency can be offered. First, realizing that it is not an easy task to create test items that effectively measure large chunks of a whole body of knowledge, the test maker often seeks a rational way to partition the body of knowledge into manageable chunks. Without such partitioning, test development would be even more difficult.

Second, most psychometricians (and other educational professionals, including teachers and textbook writers) know that to cover a large body of knowledge comprehensively, or to sample from such a body of knowledge systematically for testing purposes, it is necessary to set up a structure

for ensuring that the many varied aspects of the subject are treated. An instructional or assessment planner has to have a sense of the "parts" of the discipline that are to be covered in order to ensure that relatively even coverage is attained. Unless some sort of analytical grid or content list is used, this task is virtually impossible.

Finally, an important guiding principle in testing is the desire to provide information that can be used to help examinees, institutions, states, and other interested parties pinpoint deficiencies so as to facilitate remedial efforts. Information is the purpose and product of a test; the quality of the information is in great measure dependent on its precision. It does more good for an examinee to learn that he or she is having trouble with the passive construction and with knowledge of eighteenth-century French history than to learn that he or she is having trouble with French. Similarly, the high school that learns that, in general, its students are doing well on interpretations of literary passages but rather poorly on the comprehension of conversational French can more effectively target curricular revision than one that finds out that its students perform at the seventieth percentile in French when compared with students in all public schools in the state. The more precise information derives from the division of the field into useful groupings of content.

For the criticisms of the reductionist approach to achieve more than theoretical interest, evidence of damaging effects from the use of "atomistic" objectives and skills must be available. It may be helpful to consider an example from reading, a discipline that is probably the most

outspoken critic of testing on reductionist grounds. It has been widely asserted by advocates of the whole language philosophy of reading instruction that the division of reading, for both instructional and assessment purposes, into a supposed hierarchy of underlying or enabling skills such as letter recognition and decoding has actually done harm to those who wish to learn to read. Not only is the construct of reading violated in a conceptual sense by the assumption that skills underlie the reading act, it is contended, but the practical activity of becoming a reader is hindered or even prevented by instruction and assessment that focus on discrete skills with the assumption that practice in such skills will somehow combine to create competence in the holistic skill of reading (Goodman, 1986; Smith, 1985; Taylor, 1989). It is for this reason that the Delegates Assembly of the International Reading Association formally resolved at its May 1988 meeting that "assessment measures defining reading as a sequence of discrete skills be discouraged"; this resolution was reaffirmed in 1990.

There is considerable heated debate about this anti-skills contention, and while it may be true, it is far from proven (Carbo, 1988; Chall, 1983, 1989; McGee & Lomax, 1990; Schickedanz, 1990; Stahl, 1990; Stahl & Miller, 1989). To persuade test developers to change traditional and seemingly useful testing practices, including the division of reading into skills, it will be necessary for whole language advocates to gather evidence not only that a whole language approach "works," but also that it offers significant advantages over skills-based approaches. And even if the case is proved for reading *instruction*, it will remain to be demonstrated that reading

assessment that is conducted along skills lines is either invalid (because it does not measure what it is supposed to measure, i.e., the ability to read) or harmful (because it adversely affects instruction or learning). And, it is important to distinguish between standardized achievement tests with their emphasis on the use of normative information and which use multiple-choice test items, and multiple-choice test items and what they can measure and contribute in the way of assessment information to educators. A similar agenda of gathering research evidence is needed in other fields in which the "reductionist" approach has been criticized.

Alternative assessment approaches. Alternatives to reductionism and atomism in assessment are not clearly articulated at this time. Again, the whole language movement in language arts, which has been under development since the 1970s, presents probably the most advanced perspective. In whole language classrooms, instruction maintains a focus on reading as a whole, synthetic behavior by placing the student in an environment that is rich in print-based materials and encouraging and guiding the student to acquire, as naturally as speech is acquired, the ability to read (Goodman, et al., 1989). Assessment consists of gaining an impression over time of the student's position along the continuum of reading proficiency through repeated, diverse, and naturalistic measures (Barr, et al., 1990; Readence & Martin, 1988; Rhodes & Dudley-Marling, 1988). Such measures are many and varied. For example, observations may be taken by the teacher of the student's developing understanding of reading and of print-related concepts (e.g., the concept of a word and of a sentence, the notion that print usually proceeds from the front of a book, the top of the page, and the

left of the line) by overhearing the student reading or by considering samples of his or her writing (Glazer & Searfoss, 1988). The teacher may gain further data by questioning students orally in class. Additionally, the teacher may engage in the longitudinal collection of indirect data by encouraging students to compile work portfolios, journals, and self-written and assembled books and newspapers. For more structured assessments, the teacher may have students orally retell stories they have just read (Morrow, 1988) and may conduct miscue analyses of oral reading selections (Goodman, 1969; Goodman, et al., 1987). Such methods are an attempt to open "windows into the mind" of the reader (Goodman & Goodman, 1979), to comprehend naturalistically, gradually, and sympathetically the reader's emerging fluency in the process of reading, and to treat reading as a natural and unitary act.

How far such methods achieve their goals is a matter of debate and, more importantly, research. Is it in fact more natural for a student to compile a newspaper, enter thoughts into a journal every day, read a story out loud, or retell a story to a teacher (and often a tape recorder) (Cagney, 1988) than it is to sit before a test booklet and silently read passages and then answer sets of multiple-choice questions? And, is the gathering of miscue data while a child reads aloud in front of the teacher, which will be analyzed for the nature, quantity, and type of miscues the child has produced, any more holistic than a skills-based reading test? Is classroom observation holistic, or is it primarily feature-based, whether consciously (through the use of a structured observation form) or unconsciously (through unintentional assignment of saliency or weights to

certain characteristics rather than others)? The same questions can be asked of portfolio assessment, oral questioning, retelling analysis, and the other "holistic" methods. The need for research should be clear.

Classroom assessments, standardized achievement tests, and statewide testing programs may never move to a completely holistic approach - there is neither clear evidence that they should, nor even evidence that it is possible. But it must be noted that the critics of reductionism are having an effect. For example, serious, well-intentioned, and conscientious attempts are being made at the state level to implement assessments that at least attempt to take into account some of the more serious objections to standardized testing that have been voiced by the anti-reductionists (California Assessment Program Staff, 1989; Marzano, et al., 1988; Pikulski, 1989; Roeber & Dutcher, 1989; Valencia, et al., 1989; Wixson, et al., 1987). For instance, in the reading area, longer, unaltered passages are now being used on reading comprehension tests in place of the short, heavily edited (to meet readability levels), decontextualized snippets that were the rule until recently (Commission on Reading, National Academy of Education, 1985). College Board, too, will implement a similar change to the passages on the Scholastic Aptitude Tests. Similarly, several major test publishers will implement the same change on nationally normed standardized achievement tests.

Furthermore, the multiple-choice questions that follow these passages tend now to be more focused on comprehension than on less holistic skills. Where such lower-level skills such as decoding ability are tested, they are likely to be set in the context of authentic sentences or paragraphs.

Open-ended questions are more frequently employed now than before. Questions that focus on metacognitive strategies for reading interpretation are often included. Similar trends are going on in other subject-matter areas. While such efforts may not entirely meet the desires of critics, they substantially and undeniably represent a commitment to attend to the suggestions for change. However, even in the rush to alter traditional testing practices to attend to conscientious criticisms, there is an important need for research on whether the adjustments that are being made in the formats and contents of tests are actually effective, are free of unwanted negative consequences, and meet established and reasonable psychometric criteria for validity, reliability, and freedom from bias.

4. Under-representing the Performance of Lower-SES Examinees

Examinees from lower socioeconomic strata (SES) have typically performed less well on objective tests than have examinees of higher SES. This finding has been studied from many perspectives including one that considers the possible biases in these tests. Some advocates of authentic assessment have promoted their reforms as showing promise of addressing perceived inequities in the nature of traditional objective tests (National Commission on Testing and Public Policy, 1990; Willis, 1990).

Discussion of the criticisms. The issue of selection-type items vs. supply-type items can be viewed as a bias issue if those who construct the responses (both the "correct response" and the "distractors") are heavily but subtly influenced by their cultural understandings, and those who take the test are similarly influenced, but by different cultural understandings. What is "right" to a majority-culture test constructor may

perhaps be perceived as ridiculous, unrealistic, or just plain unthinkable to a minority-culture examinee, and what is "wrong" may appear quite reasonable. At least with a free response, it may be argued, the examinee has a chance of explaining his or her nonmainstream viewpoint.

The product vs. process controversy of objective tests (see Hambleton & Murphy, 1991) can be considered in terms of bias implications too. If the underlying reasoning or problem-solving processes of a minority-culture examinee are culture-influenced and are different from those of a majority-culture examinee, the products of the different processes may be different. In traditional product-oriented tests, there is no way for an examiner to pick up on this subtle, underlying cultural influence; at least in assessments where process is the focus, such differences in thinking and in approach to problems may be detected and either adjusted to conform to the majority process or accepted as a reasonable alternative process (with the resulting product also rethought and revalued).

A similar analysis might be applied to the higher order thinking skills issue (facts may be more salient and more highly valued in some cultures than in others; thinking ability may be surprisingly strong even in persons whose knowledge of factual content may appear deficient); the one-right-answer issue (some cultures may foster a divergent mode of thinking that regards all answers as exploratory or tentative; examinees from such backgrounds may have difficulty adopting a single-right-answer framework); the teaching-to-the-test issue (if students from low-SES backgrounds who might do better exercising their higher order thinking processes are disproportionately confined to endless reviews of facts,

their ability to enrich their thinking skills is hampered); and the reductionism issue (students from low-SES backgrounds may tend to be the ones who are stuck in the lower ends of the skills array, working on discrete skills instead of on whole, meaningful tasks) (Commission on Reading, National Academy of Education, 1985).

Alternative assessment approaches. In the previous section, various issues that are rasied when the merits of objective tests and authentic assessments are being discussed were considered from the bias perspective. This was done to show that the controversy between objective and authentic testing might have another dimension than the more frequently considered problem of the disjunction between assessment and instruction. However, the bias dimension is even less settled, if that is possible, than the other dimension, and needs research attention even more urgently, if that is possible. By no means is it clear that instituting less traditional testing methodologies will have the effect of diminishing differential performance on test items between low- and high-SES examinees. Even the moderate steps toward change in the formats of test items that are now being taken may produce ambiguous results from the perspective of item bias.

For instance, there is some evidence that African-American and female examinees may perform better on straightforward computational math items than on more "contextualized" word problems (Linn & Harnisch, 1981; Shepard et al., 1984; Scheuneman, 1987; Doolittle & Cleary, 1987; Doolittle, 1989). Will this result be reversed if the word problems are constructed to be more process-oriented than product-oriented, or more thought-provoking than

operations-intensive? Will minority examinees fare better or worse on longer reading passages bundled with several items than they have on short passages linked to one or two items? Will they do better on open-ended items than on multiple-choice ones simply because the focus and the task demands of the supply-type items are more "authentic"? Does item bias decrease when the mode of assessment is a portfolio or an observation rather than a standardized test? Will the knowledge and skills of all students, but especially of low-SES students, improve over current levels if less instructional time is devoted to preparing for fragmented, reductionist, impoverished tests? And will whatever new assessments we install achieve a more accurate appraisal of those underlying skills and abilities than traditional measures?

All these questions are empirical and should be addressed by a sound program of research. Insight into these issues should be derived from a careful and principled combination of both qualitative and quantitative research methods. Only by such a combination can the concerns and the methods of inquiry of both instructional experts and their psychometric colleagues be satisfactorily accounted for.

Conclusions

The testing field must remain alert to societal and educational demands, and, when necessary, appropriate changes must be made to testing models and practices. The availability of multiple methods of assessment in practice is generally good and useful and should be encouraged. Certainly, that was the message from Linn (1989) in the lead chapter of Educational Measurement, Haney and Madaus (1989) in their critique of

standardized achievement testing, and the developers of the teacher competencies in the area of educational testing (see, American Federal of Teachers, National Council on Measurement in Education, & National Education Association, 1990).

Objective tests, notably multiple-choice tests, have come under criticism in recent years because of concerns about their appropriateness to address current educational assessment needs. Alternatives, notably performance tests or authentic assessment, as it has been called, have promise, but they are not without their own problems. While objectivity of scoring and ease of construction and administration are not the most important criteria for determining the item format to use in a particular assessment, their importance to sound testing practice ought not to be underestimated either. For this reason, criticisms of the multiple-choice format need to be considered carefully, along with their strengths, as well as the weaknesses and strengths of any alternatives, i.e., authentic measurements. Our review in this paper of four common criticisms of objective tests suggests that the evidence against multiple-choice tests is not as strong as has been claimed; it remains to be proven that authentic measurements are always better, and that substantially more research as to the strengths and weaknesses of various item formats for meeting particular measurement needs should be carried out.

References

American Federation of Teachers, National Council on Measurement in Education, National Education Association. (1990). Standards for teacher competence in educational assessment of students. Educational Measurement: Issues and Practice, 9(4), 30-32.

Barr, R., Sadow, M. W., & Blachowicz, C. L. Z. (1990). Reading diagnosis for teachers. White Plains, NY: Longman.

Cagney, M. A. (1988). Measuring comprehension: Alternative diagnostic approaches. In S. M. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.), Reexamining reading diagnosis: New trends and procedures. Newark, DE: International Reading Association.

California Assessment Program Staff (1989). Authentic assessment in California. Educational Leadership, 46(7), 6.

Carbo, M. (1988). Debunking the great phonics myth. Phi Delta Kappan, 70, 226-240.

Carlson, D. (1991, April). Beyond the bubble: The urgency for broader assessment of school outcomes. President's address at the meeting of NCME, Chicago.

Chall, J. S. (1983). Learning to read: The great debate. New York: McGraw-Hill.

Chall, J. S. (1989). Learning to read: The great debate 20 years later- A response to "Debunking the Great Phonics Myth." Phi Delta Kappan, 70, 521-538.

Commission on Reading, National Academy of Education (1985). Becoming a nation of readers. Washington, DC: National Academy of Education.

Doolittle, A. E. (1989). Gender differences in performance on mathematics achievement items. Applied Measurement in Education, 2, 161-177.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 9, 27-32.

Glazer, S. M., & Searfoss, L. W. (1988). Reexamining reading diagnosis. In S. M. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.), Reexamining reading diagnosis: New trends and procedures. Newark, DE: International Reading Association.

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. Reading Research Quarterly, 5, 9-30.

Goodman, K. S. (1986). What's whole in whole language? Exeter, NH: Heinemann.

Goodman, K. S., & Goodman, Y. M. (1979). Learning to read is natural. In L. B. Resnick & P. A. Weaver (Eds.), Theory and practice of early reading, Vol.1. Hillsdale, NJ: Erlbaum.

Goodman, K. S., Goodman, Y. M., & Hood, W. J. (Eds.) (1989). The whole language evaluation book. Portsmouth, NH: Heinemann.

Goodman, Y. M., Watson, D. J., & Burke, C. L. (1987). Reading miscue inventory. New York: Richard C. Owen.

Hambleton, R. K., & Murphy, E. (1991). Are the criticisms of objective tests valid? (Laboratory of Psychometric and Evaluative Research Report No. 220). Amherst, MA: University of Massachusetts, School of Education.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. Phi Delta Kappan, 71(5), 683-687.

Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.), Educational measurement (3rd. ed., pp. 1-10). New York: Macmillan.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Marzano, R. J., Brandt, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., Rankin, S. C., & Suhor, C. (1988). Dimensions of thinking: A framework for curriculum and instruction. Alexandria, VA: Association for Supervision and Curriculum Development.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544.

McGee, L. M., & Lomax, R. G. (1990). On combining apples and oranges: A response to Stahl and Miller. Review of Educational Research, 60, 133-140.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? Educational Measurement: Issues and Practice, 8, 14-22.

Morrow, L. M. (1988). Retelling stories as a diagnostic tool. In S. M. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.), Reexamining reading diagnosis: New trends and procedures. Newark, DE: International Reading Association.

Nader, R. (1987). 60 years of idiocy is enough. The FairTest Examiner, 1, 1-3.

National Commission on Testing and Public Policy. (1990). From gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: National Commission on Testing and Public Policy, Boston College.

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18(9), 3-7.

Pikulski, J. J. (1989). The assessment of reading: A time for change? The Reading Teacher, 42, 80-81.

Popham, W. J. (1987). The merits of measurement-driven instruction. Phi Delta Kappan, 68, 679-682.

Readence, J. E., & Martin, M. A. (1988). Comprehension assessment: Alternatives to standardized tests. In S. M. Glazer, L. W. Searfoss, & L. M. Gentile (Eds.), Reexamining reading diagnosis: New trends and procedures. Newark, DE: International Reading Association.

Rhodes, L. K., & Dudley-Marling, C. (1988). Readers and writers with a difference. Portsmouth, NH: Heinemann.

Roeber, E., & Dutcher, P. (1989). Michigan's innovative assessment of reading. Educational Leadership, 46(7),64-69.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24, 97-118.

Schickedanz, J. A. (1990). The jury is still out on the effects of whole language and language experience approaches for beginning reading: A critique of Stahl and Miller's study. Review of Educational Research, 60, 127-131.

Shepard, L. A. (1989). Why we need better assessments. Educational Leadership, 46(7), 4-6.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Smith, F. (1985). Reading without nonsense. New York: Teachers College Press.

Stahl, S. A. (1990). Riding the pendulum: A rejoinder to Schickedanz and McGee and Lomax. Review of Educational Research, 60, 141-151.

Stahl, S. A., & Miller, P. D. (1989). Whole language and language experience approaches for beginning reading: A quantitative research synthesis. Review of Educational Research, 59, 87-116.

Taylor, D. (1989). Toward a unified theory of literacy learning and instructional practices. Phi Delta Kappan, 71, 184-193.

Valencia, S., Pearson, P. D., Peters, C. W., & Wixson, K. K. (1989). Theory and practice in statewide reading assessment: Closing the gap. Educational Leadership, 46(7), 57-63.

Wiggins, G. (1989). Teaching to the (authentic) test. Educational Leadership, 46(7), 41-47.

Wiggins, G. (1990). The case for authentic assessment. ERIC Clearinghouse on Tests, Measurement, and Evaluation. Washington, DC: American Institutes for Research.

Willis, S. (1990). Transforming the test: Experts press for new forms of student assessment. ASCD Update, 7, 3-6.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. D. (1987). New directions in reading assessment. The Reading Teacher, 40, 749-754.